Pakistan Journal of Psychology, June 2025, 56, 1, 17-25

SHIFTING PARADIGM IN PSYCHOTHERAPY: FROM HUMAN TO MACHINE: A CRITICAL REVIEW OF BENEFITS, RISK AND ETHICAL CHALLENGES

Wajeha Zainab*
Dow University of Health Sciences, Karachi, Pakistan
Zainab Khan
National University of Sciences and Technology, Islamabad, Pakistan
&
Faryal Nawab
Dow University of Health Sciences, Karachi, Pakistan

ABSTRACT

Artificial intelligence has grown markedly, transitioning from a futuristic concept in a highly technological world to something present and prevalent in everyday life. One of its most controversial applications is its emerging use as a psychotherapist. A literature search was conducted through PubMed and Google Scholar to identify qualitative and quantitative studies, as well as real-world cases, centered on the use of chatbots and conversational tools in treating psychological illnesses. This review provides a critical evaluation of the potential benefits and threats of placing AI in therapeutic roles. By examining clinical and social risks as well as ethical challenges—along with reported failures and case scenarios—the paper outlines the narrow line between assistance and harm in the digitalization of psychotherapy. The intention is not to dismiss the possible benefits of AI, but to analyze the conditions under which it may become a major threat and to highlight the safeguards required to protect patient well-being. We recommend a hybrid approach in which AI

* Correspondence Address: Wajeha Zainab, PhD; Institute of Behavioral Sciences, Dow University of Health Sciences, Karachi, Pakistan. Email: wajeha.zainab@duhs.edu.pk

functions as an assistant rather than a substitute for clinicians, supported by strict ethical principles and regulatory measures.

Keywords: *Artificial Intelligence, AI, Psychotherapy, Counseling, Mental Health*

INTRODUCTION

The question of whether a computer can think is no more interesting than the question of whether a submarine can swim." — Edsger W. Dijkstra

This is one of the questions that create an argument regarding Artificial Intelligence in psychotherapy. ELIZA and ChatGPT can be said to have evolved over time, starting as a simple program but then greatly expanding to be a conversational agent that can empathize with the user, even though only to a certain extent. They appear when the whole world is facing a mental-health care shortage. World Health Organization (2025) reports that more than 75% of the world population in low-and middle-income nations do not access mental health care. Thus, human and financial constraints along with stigma spark curiosity in AI as a possible solution to fulfill the unmet needs.

In 1966, Weizenbaum developed ELIZA, the first computer-based counseling in a client-based perspective. Later decades were marked by appearance of therapy chatbots applying CBT, Mindfulness and Acceptance and Commitment Therapy (e.g. Woebot, Wysa, Replika, XiaoNan). In recent years, AI has been popularized in the area of emotional support due to the recent explosion of large language models (LLMs) and especially ChatGPT due to its free offering and advanced natural-language processing (Zao-Sanders, 2025). The ChatGPT-4 has been shown to have advanced reasoning and theory of mind capabilities, with reasoning capacity approximating that of adults' social cognition and reasoning (Cheng et al., 2023; Krach et al., 2008). It is even able to generate contextually emotionally relevant responses (Amin et al., 2023), which is the source of optimism about its potential application in therapy (Cross et al., 2024; Khawaja & Bélisle-Pipon, 2023). In addition, greater capabilities are likely to be developed in open AI models that widen the scope of services and its competence in mental health care (Priyanka et al., 2024).

Therefore, this review addresses the question whether the artificial intelligence models can be an effective replacement for human therapists? More

specifically, it examines the potential benefits, clinical, and social risks as well as ethical challenges of AI therapists to the patients' well-being.

METHOD

Literature Search Strategy

The literature search was conducted using the PubMed and Google Scholar databases. Keywords such as *"Artificial Intelligence," "Psychotherapy," "AI chatbots," "Mental Health,"* and *"AI Therapist"* were used to identify relevant studies. No restriction was applied to the publication time frame.

Study Selection Process

As this review specifically focused on the use of AI as a psychotherapist, studies were included if they met the following criteria:

- Qualitative or quantitative studies published in peer-reviewed journals, conference papers, book chapters, review papers or academic reports.
- Publications written in English with full-text availability.
- Studies involving human participants of any age group, as well as clinicians or therapists utilizing AI-based psychotherapy tools to improve mental health.
- Studies examining AI chatbots (e.g., Woebot, Wysa), conversational agents, digital CBT platforms, or machine-learning-driven therapy applications.

CRITIQUE AND DISCUSSION

Benefits of AI in Clinical Practice

There are many advantages of AI chatbots: they can be deployed on a massive scale, accessible by users of all geographical and income bases, and 24/7. They are able to offer psychoeducation, cognitive restructuring (Wang et al., 2025) and behavioral activation (Jia, 2025) and minimize the fear of judgment (Chaudhry & Debi, 2024; Sundar & Kim, 2019; Ta et al., 2020). AI can be a non-judgmental listener to the adolescents who are at times sensitive to being misunderstood. To individuals who are reluctant to pursue formal treatment, AI can act as a starting point to the assistance, without any stigma.

One meta-analysis of randomized controlled trials (RCTs) that provide interventions, mostly relying on the cognitive-behavioral approach, and, in certain instances, positive psychology or mindfulness, discovered that AI-based chatbots demonstrate good potential efficiency in reducing depression and anxiety symptoms in adults (Zhong et al., 2024). Nevertheless, the fact that current therapy chatbots are generally based on manualized treatment plans is one significant limitation to the current research on this field. Popular open-domain models (including ChatGPT and similar AI systems) are not necessarily thoroughly tested in terms of their effectiveness in addressing mental health problems, despite being used by more people every day.

Clinical Risks and Ethical Challenges

Similar to all innovations, AI psychotherapists have their positive and negative sides. Safety is one of the issues of concern. AI-based chatbots do not meet the necessary criteria to handle the rising suicidal threat. In a lawsuit in 2024, it was alleged that an AI chatbot had provoked suicidal ideation in a 14-year-old (CourtListener, 2024) and a second lawsuit alleged it had contributed to the suicide of a 13-year-old.

There is also the tendency of AI agents to authenticate maladaptive, bizarre, and dangerous beliefs of users rather than criticize them. This is associated with their system of user satisfaction, sometimes called AI sycophancy, in which user engagement is maximized (Sharma et al., 2023). This reinforcement may cause dependency on AI agents (Liu & Sundar, 2018; Skjuve et al., 2021; Ta et al., 2020) which result in the development of pathological patterns, including social isolation, less socialization, underdeveloped socialization skills, and the inability to accept a different point of view (Fang et al., 2025).

Moreover, AI-based systems do not have moral attention, subtlety, and responsibility. Hipgrave and his colleagues (2025) have performed a qualitative study where a clinician has made their concerns recorded as: It may not be capable of comprehending some of the very nature of subtleties of human type of feelings […] At times we get some patients who present in treatment some 10 sessions, and they simply state, I am fine, I am fine. There's nothing wrong with me. And it requires approximately 10 sessions to become acquainted with what is the problem. Therefore, I do not understand how far AI can go in such situations

when the patient declares that he is fine, but in reality, the situation is not the same.

In a separate study, a review of negative user response to the conversational chatbot Replika identified 800 cases of interest, out of a total 35,105 reviews, where users complained of unsolicited sexual advances, unwanted continued unwanted inappropriate interactions to high degree, and the lack of respect to personal boundaries by the chatbot. Lots of users reported that they felt uncomfortable, their privacy was intruded upon, and they were disappointed, particularly the ones who wanted a platonic or therapeutic AI companion (Sharma et al., 2023). An app that was initially created as a "safe space" where people can express their opinions became a traumatic experience to some users, including minors. This brings the very vital question: *who bears the responsibility since machines have no intentions of a human being?*

Privacy breach, ambiguity of who owns the data and commercialization of user vulnerability are also considered to be ethical hazards. An example is that the personal information obtained during the sessions can be recorded on cookies and used by algorithms to sell specific products, such as psychiatric medication, created in accordance with the vulnerabilities of the user (Williams et al., 2025). Similar content on other social media platforms further promoted similar ideas regardless of the extent of risk this may be posing.

Lastly, the excessive dependence on chatbots may delay access to the immediate professional care. The goal of human therapists is to ensure that clients learn to solve problems and become resilient in order to avoid a dependency in the future. Chatbots, on the contrary, tend to offer pre-coded solutions, which might not only reduce actual skill growth, but also create dependency on the system (Hipgrave et al., 2025).

Social Risks: Erosion of Therapeutic Alliances and Human Connections

Jonathan Shedler's assertion that "the therapeutic relationship is not just a container for treatment; it is the treatment" captures the essence of psychotherapy. The primary change agent is the alliance which is founded on trust, empathy and shared meaning. The human mind combines emotion, body language, verbal expression, history, present conditions, and ethical accountability whereas machines only produce statistically viable answers but not meaningful ones. Moreover, even the therapist-client relationship in itself can

also be utilized to reveal the patterns of the client by the process of transference. Even though certain conversational agents and chatbots are designed to be emotionally intelligent (Ghandeharioun et al., 2019; Mumuksh et al., 2020) and commit to establish a therapeutic relationship with users (Darcy et al., 2021), AI cannot help one feel warm unless it provides unconditional help. It is unable to participate in genuine emotional attentiveness and be ethically responsible, which puts inherent boundaries on its work as a psychotherapist.

Though AI chatbots are associated with some risks, there is no denying the benefits of AI chatbots. To obtain the maximum opportunities and reduce the possible harms, a balanced strategy should be developed to incorporate AI into the psychotherapy process. Instead of providing AI with the driving seat, a hybrid approach might prove to be more efficient, and the AI could be used as a complement to human care. AI is able to assist medical workers in diagnosis and in the development of individual treatment projects. It is also able to improve the decision-making process of clinicians during psychotherapy (Plakun, 2023). Also, this underscores the necessity of a research on the potential benefits of applying AI in psychotherapy to clients and therapists.

Future Directions

There are a number of randomized controlled trials examining the effectiveness of AI applications based on manualized treatments; however, there is a lack of evidence on the effectiveness of large language models, such as ChatGPT, in managing mental health–related issues. Large-scale randomized controlled trials are needed to generate stronger evidence on the clinical effectiveness and long-term therapeutic outcomes of AI-based psychotherapy. Furthermore, future research should test hybrid models that integrate human therapists with AI systems, combining the emotional intelligence of clinicians with the scalability and technological strengths of AI. Studies are also required to explore the capabilities of AI in early intervention, such as detecting early signs of crisis situations including self-harm and suicidal ideation and to determine how system can be established to inculcate in emergency referral pathways.

Policy Implications

The incorporation of AI into clinical practice requires strong policy implementation and regulation. Mental health authorities and regulatory bodies must approve AI-based mental health tools to ensure ethical standards, data

security, and system efficiency in managing crisis situations. Ethical frameworks should include mandatory supervision by mental health professionals, clearly defined scope and limitations of the AI tool, crisis-response mechanisms, and transparency in data ownership. Regulatory bodies should also conduct periodic audits to monitor safety standards and evaluate outcomes.

Conclusion

The development of AI as a form of psychotherapeutic agent has both potential benefits and threat. It may be useful in filling the mental health treatment gap worldwide, but its potential to replace human therapists is still not as high as its availability and wide applicability; clinical and ethical issues prevail at the moment. Continuous advancements of technology and AI models provide a way towards the future, where AI is used as a complementary tool that would optimize the benefits and would not be used as a total replacement of mental health care.

REFERENCES

Amin, M. M., Cambria, E., & Schuller, B. W. (2023). Will affective computing emerge from foundation models and general artificial intelligence? A first evaluation of ChatGPT. *IEEE Intelligent Systems*, *38*(2), 15-23.

Chaudhry, B. M., & Debi, H. R. (2024). User perceptions and experiences of an AI-driven conversational agent for mental health support. *Mhealth*, *10*, 22.

Cheng, S. W., Chang, C. W., Chang, W. J., Wang, H. W., Liang, C. S., Kishimoto, T., ... & Su, K. P. (2023). The now and future of ChatGPT and GPT in psychiatry. *Psychiatry and Clinical Neurosciences*, *77*(11), 592-596.

CourtListener. (2024). Garcia v. Character Technologies, Inc., No. 6:24-cv-01903 (M.D. Fla., filed Oct. 22, 2024). https://www.courtlistener.com/docket/69300919/garcia-v-character-technologies-inc/

Cross, S., Bell, I., Nicholas, J., Valentine, L., Mangelsdorf, S., Baker, S., Titov, N., & Alvarez-Jimenez, M. (2024). Use of AI in Mental Health Care:

Community and Mental Health Professionals Survey. *JMIR Mental Health, 11*, e60589. https://doi.org/10.2196/60589.

Fang, C. M., Liu, A. R., Danry, V., Lee, E., Chan, S. W., Pataranutaporn, P., ... & Agarwal, S. (2025). How AI and human behaviors shape psychosocial effects of chatbot use: A longitudinal randomized controlled study. *arXiv preprint arXiv:2503.17473*.

Hipgrave, L., Goldie, J., Dennis, S., & Coleman, A. (2025). Balancing risks and benefits: clinicians' perspectives on the use of generative AI chatbots in mental healthcare. *Frontiers in Digital Health*, *7*, 1606291.

Jia, E., Macon, J., Doering, M., & Abraham, J. (2025). Effectiveness of Digital Behavioral Activation Interventions for Depression and Anxiety: Systematic Review and Meta-Analysis. *Journal of Medical Internet Research*, *27*, e68054.

Khawaja, Z., & Bélisle-Pipon, J. C. (2023). Your robot therapist is not your therapist: understanding the role of AI-powered mental health chatbots. *Frontiers in Digital Health*, *5*, 1278186.

Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., & Kircher, T. (2008). Can machines think? Interaction and perspective taking with robots investigated via fMRI. *PloS One*, *3*(7), e2597.

Liu B and Sundar SS (2018) Should machines express sympathy and empathy? Experiments with a health advice chatbot. *Cyberpsychology, Behavior, and Social Networking 21*(10), 625–636.

Plakun, E. M. (2023). Psychotherapy and artificial intelligence. *Journal of Psychiatric Practice®*, *29*(6), 476-479.

Priyanka, Kumari, R., Bansal, P., & Dev, A. (2024). Evolution of ChatGPT and different language models: A review. In *International Conference on Smart Computing and Communication* (pp. 87-97). Singapore: Springer Nature Singapore.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., ... & Perez, E. (2023). Towards understanding sycophancy in language models. *Arxiv Preprint Arxiv:2310.13548*.

Skjuve, M., Følstad, A., Fostervold, K. I., & Brandtzaeg, P. B. (2021). My Chatbot companion: A study of human-chatbot relationships. *International Journal of Human-Computer Studies 149,* 102601.

Sundar, S. S., & Kim, J. (2019). Machine heuristic: When we trust computers more than humans with our personal information. In *Proceedings of the 2019 CHI Conference on human factors in computing systems* (pp. 1-9).

Ta, V., Griffith, C., Boatfield, C., Wang, X., Civitello, M., Bader, H., ... & Loggarakis, A. (2020). User experiences of social support from companion chatbots in everyday contexts: thematic analysis. *Journal of Medical Internet Research*, *22*(3), e16235.

Wang, Y., Li, X., Zhang, Q., Yeung, D., & Wu, Y. (2025). Effect of a Cognitive Behavioral Therapy–Based AI Chatbot on depression and loneliness in Chinese University Students: Randomized controlled trial with financial stress moderation. *JMIR mHealth and uHealth*, *13*, e63806.

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36-45.

Zao-Sanders, M. (2025). How people are really using generative AI now. *Harvard Business Review*. https://hbr.org/2024/03/how-people-are-really-using-genai

Zhong, W., Luo, J., & Zhang, H. (2024). The therapeutic effectiveness of artificial intelligence-based chatbots in alleviation of depressive and anxiety symptoms in short-course treatments: A systematic review and meta-analysis. *Journal of Affective Disorders, 356,* 459-4694